Lagrange Fase 2: un supercalcolatore da Top500 per la ricerca in Lombardia

Claudio Arlandini
CILEA, Segrate

Abstract
Nell'ambito dell'Iniziativa LISA si è provveduto all'installazione di una piattaforma hardware di nuova generazione, concepita come espansione del cluster *lagrange*. Vengono presentati una descrizione dettagliata delle caratteristiche dell'espansione e i primi risultati riguardanti le prestazioni. Il sistema nella sua nuova veste è stato classificato alla 210° posizione della classifica TOP500 di giugno 2010.

In the framework of the LISA Initiative a new hardware platform was installed at CILEA, planned as expansion of the *lagrange* cluster. This article presents a detailed description of the expansion and the first performance results. The system in its complete form was listed at rank 210 of the TOP500 list of June 2010.

Keywords: Iniziativa LISA, Supercalcolo, TOP500.

Introduzione

Gli scopi dell'Iniziativa LISA vedono la necessità di una piattaforma hardware all'avanguardia da mettere a disposizione della ricerca lombarda, che possa configurarsi come competitiva sul piano prestazionale a livello europeo. Il Comitato di Indirizzo, sentito il parere dei tecnici, ha stabilito che l'acquisto andasse configurato come espansione del sistema esistente piuttosto che come un nuovo calcolatore a se stante, al fine di ottimizzare il rapporto costo/prestazioni. L'architettura di *lagrange* infatti era già stata ingegnerizzata con un'ottica di espandibilità, e, non meno importante, il sistema ha dato prova di una notevole affidabilità nel corso del suo servizio. Sono stati quindi scelti due fornitori, Hewlett Packard (HP) e Dell, ai quali è stato chiesto di fornire una adeguata offerta tecnologica per l'espansione. Dopo un'accurata analisi, la proposta HP è stata giudicata vincente.

Struttura dell'espansione

Fig. 1 – Una visione d'insieme di lagrange nella sua nuova configurazione

La soluzione adottata (Fig. 1) è composta da Server HP BladeSystem basati su processore Intel di ultima generazione. I server Blade sono di modello BL2x220c [1], alloggiati in BladeSystem c7000 [2], enclosure da 10U dotati di 16 baie e che consolidano al loro interno alimentazione, raffreddamento, connettività e ridondanza.

Fig. 2 – Dettaglio sulle lame BL2x220c

Le lame BL2x220c (Fig. 2) sono all'avanguardia per quanto riguarda la compattezza, riunendo in un unico server due motherboard indipendenti con due socket ciascuno. Questo, utilizzando i nuovi processori exacore, permette di triplicare il numero di core per enclosure rispetto al *lagrange* originario. Le componenti all'interno di un'enclosure sono connessi mediante un Signal Midplane per il trasporto dei segnali a 5 Terabit/s, totalmente passivo e ad alta affidabilità, e ad un Power Backplane per la distribuzione dell'alimentazione. Il sistema di alimentazione è composto da 6 power supplies sui quali è possibile definire diversi schemi di ridondanza, mentre il raffreddamento è ottenuto mediante 10 ventole. Ogni componente dell'enclosure è hot-pluggable per garantire facilità di intervento e quindi il minimo di downtime in caso di guasti. Lo switch Infiniband (IB) interno all'enclosure è di tecnologia Quad Data Rate (QDR) a 40 Gb/s, mentre enclosures diverse sono collegate tra loro tramite il preesistente switch Double Data Rate (DDR). Questa scelta consente un notevole risparmio a scapito di una modesta perdita di prestazioni. Infatti, data l'elevata densità del sistema, ogni enclosure contiene ben 384 core computazionali. Questo significa che, grazie anche ad un sapiente tuning del sistema di code e load balancing (PBS Professional), solo simulazioni richiedenti un così elevato numero di threads o superiore saranno costretti a ricorrere alla rete esterna DDR per il message passing.

In totale l'espansione a disposizione dell'Iniziativa LISA si compone di 4 siffatte enclosure, corrispondenti a 64 lame BL2x220c, ovvero 128 nodi di calcolo, configurati ciascuno con:

- 2 CPU Intel Xeon X5660
- 24 GB RAM
- 1 disco 120GB SATA

L'aggiunta di un numero così elevato di nodi di calcolo ha reso necessaria anche la ristrutturazione dell'architettura del sistema stesso e dello storage principale. Le prestazioni dello storage HP MSA1500 si erano infatti rivelate il collo di bottiglia principale del cluster nella sua prima versione. In accordo con gli esperti di HP si è quindi deciso di sostituire il sistema MSA1500 con uno storage HP EVA4400 [3] della capacità di 9 TB lordi, dalle prestazioni più elevate. Inoltre, anche per ragioni di incremento della stabilità e del livello di sicurezza, si è deciso di separare a livello hardware le funzioni di accesso al sistema da parte degli utenti da quelle di file-serving, ora ottenuto mediante sistemi ridondati connessi allo storage con connessioni a fibre ottiche FibreChannel, e da alcune funzioni fondamentali di gestione del cluster.

HP ha inoltre offerto un server Proliant SL6500, un prodotto appena uscito sul mercato costruito per ospitare gli acceleratori GPU di ultima generazione NVIDIA M2050 [4] (Fig. 3).

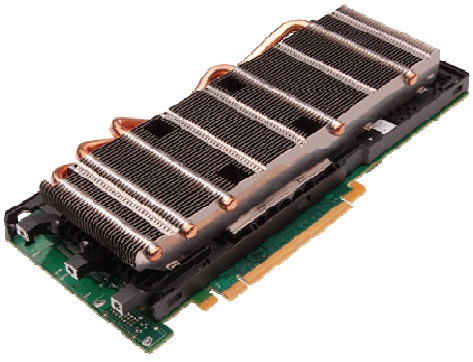

Fig. 3 – La scheda NVIDIA M2050

Basati sull'architettura CUDA™ di nuova generazione dal nome in codice "Fermi", i moduli di GPU Computing Tesla M2050 permettono un'integrazione perfetta del GPU Computing con i sistemi host per l'elaborazione a elevate prestazioni. Le GPU Tesla serie 20 sono le prime a fornire più di dieci volte la potenza in doppia precisione di una CPU x86 quad-core e le prime ad avere la memoria ECC. Il server SL6500 è un sistema che offre connettività Ethernet 10Gb e IB QDR sulla motherboard e che può ospitare fino a 8 GPU. Questo sistema sarà disponibile agli utenti a partire dalla fine del mese di ottobre e costituirà un fondamentale strumento di sviluppo per alcuni dei progetti dell'Iniziativa LISA..

Analisi preliminare delle prestazioni

Il processore di cui sono dotati i nodi di calcolo appartiene all'ultima generazione di processori Intel, ovvero la serie 5600 denominata in codice "Westmere". Il processore X5660 [5] ha una frequenza di 2.80 GHz, 6 core computazionali, una cache di livello 3 di 12 MB, un consumo di 95W e supporta le nuove tecnologie denominate Turbo Boost [6] e Hyper-Threading [7].

La tecnologia Turbo Boost permette di incrementare le prestazioni aumentando la frequenza operativa dei core computazionali quando le condizioni operative lo consentono, tipicamente quando di una CPU si utilizza un numero di core minore rispetto a quelli disponibili. L'Hyper-Threading permette invece di gestire simultaneamente due threads per core computazionale, ottimizzando così tutti i cicli di clock. Non è una tecnologia nuova, era già stata introdotta in processori di generazioni precedenti, ma nella serie 5600 è stata notevolmente migliorata nelle prestazioni. Al contrario del Turbo Boost, è un'opzione che va attivata esplicitamente al boot del nodo. Al momento non è stata attivata di default su *lagrange*, in attesa del completamento dei test prestazionali in corso.

Inoltre, per la comunicazione tra processori e le rispettive RAM viene supportata la tecnologia QuickPath che consente connessioni point-to-point all'interno del nodo di 25,6 GB/s, molto superiore rispetto alle generazioni precedenti. Questo significa un miglior accesso alla memoria e quindi un ulteriore incremento delle prestazioni delle applicazioni memory-bound.

Ultima innovazione, ma non meno importante, all'interno del server è attiva la tecnologia Intel Intelligent Power [8] che riduce drasticamente i consumi elettrici di componenti del server quando non vengono utilizzati, riducendo i consumi di server non occupati in simulazioni fino al 50%.

Tutte queste nuove tecnologie si traducono in un aumento delle prestazioni a livello di applicazione. La quantificazione di questo aumento dipende chiaramente dalla tipologia della stessa, ma i primi test effettuati mostrano trattarsi mediamente di un fattore 2 rispetto ai vecchi processori di *lagrange*, che pure avevano una frequenza superiore.

I nostri test sono ancora in corso. Giusto per citare i primi risultati, ecco le prestazioni ottenute per quanto riguarda l'applicazione di analisi strutturale Simulia Abaqus [9], versione 6.7-3, sul benchmark standard E1 [10]. Il benchmark consiste nella simulazione di un automobile che impatta un muro rigido alla velocità di 25 mph. La geometria è relativamente piccola, con ca. 280.000 elementi.

Il grafico (Fig. 4) evidenzia un aumento delle prestazioni rispetto alla generazione precedente di almeno un 30%, ma anche un sensibile miglioramento della scalabilità.

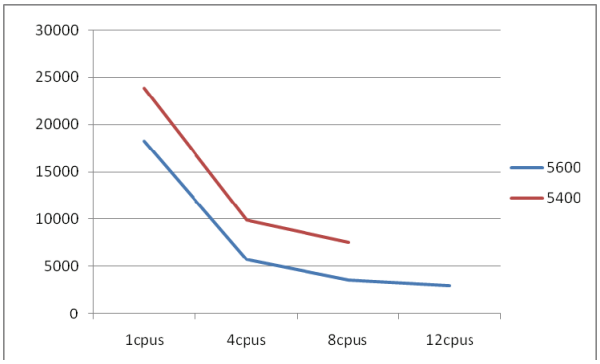

Fig. 4 – Tempi di calcolo (in s) per il Benchmark E1 di Abaqus sui nodi di lagrange vecchia versione (CPU serie 5400) e nuova (CPU serie 5600).

La classifica TOP500

Tra le intenzioni per le quali l'Iniziativa LISA è stata avviata era fondamentale il mettere a disposizione della ricerca lombarda una infrastruttura hardware competitiva a livello europeo. La classifica TOP500 [11] di giugno, che vede *lagrange* nella sua forma espansa alla 210° posizione (64° tra i sistemi europei), dimostra che questo obiettivo è stato raggiunto.

La classifica TOP500 elenca semestralmente i 500 calcolatori più potenti al mondo. La potenza viene calcolata attraverso un benchmark

standard LINPACK, la soluzione di un sistema denso di equazioni lineari [12].
Le misurazioni effettuate hanno dato per il sistema completo una potenza reale di 35.7 TFlop/s rispetto ai 45.2 TFlop/s di picco, con un'efficienza prossima al 79%.

Il CILEA appare per la quattordicesima volta nella graduatoria, pubblicata nella prima volta nel 1993. La posizione migliore mai ottenuta è stata la 135°, nel giugno del 2008, grazie alla prima installazione di *lagrange*.

L'iniziativa LISA appare quindi partita sotto i migliori auspici. Non resta che augurarsi che risultati scientifici di buon livello premino gli sforzi di tutti.

Bibliografia

[1] URL: http://h10010.www1.hp.com/wwpc/us/en/en/WF05a/3709945-3709945-3328410-241641-3722790-3707371.html
[2] URL: http://h18000.www1.hp.com/products/blades/components/enclosures/c-class/c7000/
[3] URL: http://h10010.www1.hp.com/wwpc/us/en/en/WF05a/12169-304616-304648-304648-304648-3664763.html
[4] URL: http://www.nvidia.com/object/product_tesla_M2050_M2070_us.html
[5] URL: http://ark.intel.com/Product.aspx?id=47921
[6] URL: http://www.intel.com/technology/turboboost/index.htm
[7] URL: http://www.intel.com/technology/platform-technology/hyper-threading/index.htm
[8] URL: http://www.intel.com/technology/intelligentpower/index.htm
[9] URL: http://www.simulia.com/
[10] URL: http://www.simulia.com/support/v67/v67_performance.html#XplServerData
[11] URL: http://www.top500.org/
[12] URL: http://www.netlib.org/benchmark/hpl/